Comment | Vol. 10, No. 9 | 1 Sep 2019 | BIOMEDICAL OPTICS EXPRESS 4489
Biomedical Optics EXPRESS

Use of Raman spectroscopy to screen diabetes mellitus with machine learning tools: comment

IVAN A. BRATCHENKO,* DMITRY N. ARTEMYEV, YULIA A. KHRISTOFOROVA, AND LYUDMILA A. BRATCHENKO

Laser and Biotechnical Systems Department, Samara National Research University, Moskovskoe shosse 34, Samara 443086, Russia
**iabratchenko@gmail.com*

Abstract: This paper comments on the article "Use of Raman spectroscopy to screen diabetes mellitus with machine learning tools" by E. Guevara et al. The authors propose an optical method for noninvasive automated screening of type 2 diabetes mellitus. Despite the high performance of the proposed method, results shown by the authors may be ambiguous due to the overestimation of classification models for Raman spectral data analysis.

In their recent publication on Raman spectroscopy application for the non-invasive measurements of human skin optical properties in order to determine the presence of type 2 diabetes mellitus (DM), E. Guevara et al. propose using the Raman spectroscopy system coupled with several supervised machine-learning techniques to discern between DM patients and healthy controls. The application of artificial neural networks (ANN) and principal component analysis (PCA) was carried out to perform discrimination between the DM and control groups [1]. Despite the high performance of the proposed method, the results shown by the authors could be treated incorrectly because of the lack of statistical data description and possible overestimation of models for Raman data analysis.

The authors used a Raman system with a 785 nm central wavelength and focused a 90 mW beam into a spot of 200 μm. The acquisition time was 15 s. According to ANSI or similar standards, there is a limit of 1.63 W/cm^2 with 785 nm laser source irradiance of skin with more than 10 s exposure times [2]. Utilized settings lead to the 287 W/cm^2 irradiance of the estimated skin sample. Most likely, the authors missed some details about the procedure of spectra registration, as accurate estimation of skin irradiation can be challenging due to focusing and dispersing of light at the distal tip of the probe. The authors need to calculate correctly the intensity density on the skin to find if the proposed approach satisfies the safety standards.

Regardless to the spectra registration procedure, there may be several drawbacks in the statistical data analysis connected with over-estimation of utilized classifiers and the number of analyzed spectra. For the analysis of spectral data from 20 people (11 with DM and 9 healthy subjects, 5 spectra were registered for each tested tissue sample) authors used a feed-forward ANN classifier [3] and a support vector machine (SVM) classifier [4] for the principal components (PC) obtained during the PCA [5]. However, the number of spectra utilized for the analysis by E. Guevara et al. is not clear. For example, Fig. 3 (A and C) in the original article demonstrates ROC curves of the proposed classifiers, and ROC curves of the ANN classifier look smoother than ROC curves of the SVM classifier. It appears that E. Guevara et al. utilized 100 spectra (20 subjects x 5 scans at each point) in the ANN analysis and 20 spectra (9 healthy + 11 DM) in the SVM analysis. Moreover, Fig. 5 in the commented article contains only 20 points, which indicates that E. Guevara et al. utilized 20 spectra in the SVM analysis.

#352608 https://doi.org/10.1364/BOE.10.004489
Journal © 2019 Received 23 Nov 2018; revised 5 Feb 2019; accepted 6 Feb 2019; published 6 Aug 2019

Biomedical Optics EXPRESS

In their publication, E. Guevara et al. implemented ANN with a single-hidden-layer with 14 neurons and one output neuron. The number of neurons was selected using the method devised by Huang [6]. However, Huang proposed to use a two-hidden-layer feedforward network with n hidden neurons to distinct N samples with any arbitrarily small error (assuming $n<<N$). Therefore using 14 neurons to classify 20 objects (20 or 100 spectra) in the data set seems questionable. Note, that ANN may be so powerful that it is able to obtain a nonlinear function that reconstructs in detail the values of the targets starting from the data about inputs if every input-target pair is known to the ANN. But in this case, ANNs may overfit data and no realistic regression law can be obtained, thus, it is proper to consider only a small number of hidden neurons in case of small data sets analysis [7]. In fact, the number of hidden layers and their respective amount of neurons depend on the nature and complexity of the problem being mapped by the ANN, as well as the quantity and quality of the available data about the problem [3]. Thus, in order to provide a reliable data, the authors should test a number of ANN variations (or use regularization in proposed ANN) to ensure that the obtained performance of ANN is not an accidental result caused by overfitting.

E. Guevara et al. utilized 14 – 15 PCs using the Bartlett's chi-square test in PCA. Although the Bartlett's test is a strict mathematical criterion, the PCs appropriate for the analysis should be also determined by means of certain data set features [8]. Raman spectral data set analysis may involve estimating the shape of PCs (or loadings) to make sure that PCs contain useful information about the chemical composition of the tested tissue.

E. Guevara et al. tried to validate the obtained ANN and PCA-SVM classifiers by implementing 10-fold cross-validation (CV). CV is a well-known tool for the verification of the classifier performance, but a serious over-estimation is more likely to occur if the validation data set bear significant variances compared with the training data set [9]. As the authors used a small data set, there is a great chance that validation and training sub-sets could differ dramatically, and, thus, classifying models described by the authors may be over-estimated. Moreover, utilization of the CV procedure by E. Guevara et al. is not clearly described, and most likely E. Guevara et al. performed PCA on the entire data set prior to CV (that may be confirmed by reference to the study of Dingari et al. [10] in the commented article). In contrast to Dingari et al. where 4 PCs capture 99.7% of the total variation of their data, 15 PCs in commented article capture only 95% of the total data set variation, and, thus, performing PCA prior to CV may be incorrect [11] (i.e. keeping the PCA outside CV loop, E. Guevara et al. are ignoring a potentially significant source of variation).

The authors estimated the Pearson's correlations between the Raman spectra of the advanced glycation end products (AGE) (accumulating in skin during DM progression) and obtained PCs. For the inner arm data set, almost all calculated values of the Pearson's correlations are 0, which may indicate that the chosen PCs may be insufficient for DM identification. However, the authors state that "Using ANN, the skin location with the highest classification accuracy is the inner arm". Thus, the authors should provide some explanation of such contradictory findings (i.e. provide data about correlations of raw Raman spectra of different skin sites and AGE).

It is important to note, that E. Guevara et al. (1) could have used less complex classification models (e.g. make the test sets twice as large, replace the SVMs by the logistic regression classifier, consider (random) subsets of 2, 3, …, 14 out of 15 PCs, in addition to all 15, and measure their performance recomputing PCA inside the training sets to see if that makes a difference); (2) may try to simplify representation of classifiers accuracy (Fig. 3 B and D), as reporting so many different measures of performance is not really helpful; it may be much more valuable to fix, for example, the sensitivity of the classification scheme at 95%, and therefore the researcher goal would be to advance their methods by improving the corresponding specificity.

In summary, to clear away the ambiguity due to the limited size of the data set and to avoid overestimation of the proposed classification models the study by E. Guevara et al.

Biomedical Optics EXPRESS

needs additional data regarding the utilized number of analyzed spectra, PCs shape, ANN architecture, CV procedure, etc.

Acknowledgements

We are grateful to the reviewers for critical comments.

Disclosures

The authors declare that there are no conflicts of interest related to this article.

References

1. E. Guevara, J. C. Torres-Galván, M. G. Ramírez-Elías, C. Luevano-Contreras, and F. J. González, "Use of Raman spectroscopy to screen diabetes mellitus with machine learning tools," Biomed. Opt. Express **9**(10), 4998–5010 (2018).
2. N. Kourkoumelis, I. Balatsoukas, V. Moulia, A. Elka, G. Gaitanis, and I. D. Bassukas, "Advances in the in Vivo Raman Spectroscopy of Malignant Skin Tumors Using Portable Instrumentation," Int. J. Mol. Sci. **16**(7), 14554–14570 (2015).
3. I. N. da Silva, *Artificial Neural Networks, A Practical Course* (Springer 2017).
4. C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn. **20**(3), 273–297 (1995).
5. G. Hughes, "On the mean accuracy of statistical pattern recognizers," IEEE Trans. Inf. Theory **14**(1), 55–63 (1968).
6. G.-B. Huang, "Learning capability and storage capacity of two-hidden-layer feedforward networks," IEEE Trans. Neural Netw. **14**(2), 274–281 (2003).
7. A. Pasini, "Artificial neural networks for small dataset analysis," J. Thorac. Dis. **7**(5), 953–960 (2015).
8. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. (Wiley, 2001).
9. S. Guo, T. Bocklitz, U. Neugebauer, and J. Popp, "Common Mistakes in Cross-Validating Classification Models," Anal. Methods **9**(30), 4410–4417 (2017).
10. N. C. Dingari, G. L. Horowitz, J. W. Kang, R. R. Dasari, and I. Barman, "Raman spectroscopy provides a powerful diagnostic tool for accurate determination of albumin glycation," PLoS One **7**(2), e32406 (2012).
11. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction,* 2nd edition (Springer, 2008).